AFRL-HE-BR-TR-2004-0056

United States Air Force
Research Laboratory

THE SENSITIVITY AND SPECIFICITY OF OCULOMETRICS UNDER FATIGUE STRESS, COMPARED TO PERFORMANCE AND SUBJECTIVE MEASURES

James C. Miller

HUMAN EFFECTIVENESS DIRECTORATE
BIOSCIENCES AND PROTECTION DIVISION
FATIGUE COUNTERMEASURES BRANCH
2485 GILLINGHAM DRIVE
BROOKS CITY-BASE, TX 78235

Douglas R. Eddy

NTI Inc.
8248 CHENNAULT DRIVE
BROOKS CITY-BASE, TX 78235

Joseph Fischer

GENERAL DYNAMICS
P.O. BOX35482
BROOKS CITY-BASE, TX 78235

May 2004

NOTICES

This report has been reviewed and is approved for publication.

//SIGNED//

JAMES C. MILLER, Ph.D.
Project Scientist

//SIGNED//

F. WESLEY BAUMGARDNER, Ph.D.
Deputy, Biosciences and Protection Division

DESTRUCTION NOTICE: Destroy by any method that will prevent disclosure of contents or reconstruction of the document.

WARNING-This document contains technical data whose export is restricted by the Arms Export Control Act (22 U.S.C. 2751, et seq.) or the Export Administration Act of 1979, as amended (Title 50 U.S.C. 2401, et seq.) Violating these export-control laws is subject to severe criminal penalties. Dissemination of this document is controlled under DoD Directive 5230.25.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
May 2004

2. REPORT TYPE
Interim

3. DATES COVERED *(From - To)*
Oct 2002-Mar 2004

4. TITLE AND SUBTITLE
The Sensitivity and Specificity of Oculometrics under Fatigue Stress, Compared to Performance and Subjective Measures

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)
James C. Miller, Douglas R. Eddy, Joseph Fischer

5d. PROJECT NUMBER
7757

5e. TASK NUMBER
P9

5f. WORK UNIT NUMBER
05

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NTI Inc.
8248 Chennault Drive
Brooks City-Base, TX 78235

General Dynamics
P. O. Box 35482
Brooks City-Base, TX 78235

8. PERFORMING ORGANIZATION REPORT

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Effectiveness Directorate
Biodynamics and Protection Division
Fatigue Countermeasures Branch
2485Gillingham Drive
Brooks City-Base, TX 78235

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HE

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-BR-TR-2004-0056

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release, distribution unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We wished to compare the sensitivity of oculometric measures under fatigue stress to the sensitivities of performance and subjective measures, and to assess the specificity of oculometrics for predicting performance effects . We used data from the first night of sleep deprivation in a sleep aids study. Each of 13 subjects was represented four times in the final data set, with each of these four nights separated by at least a week. We focused on oculometric, simple cognitive task and subjective data and on two test periods within the night of sleep deprivation: at 21:00 and 03:00. All data were standardized as within-subject z scores across weeks. There were large, reliable differences in cognitive task performance and subjective assessments, in the expected directions, across the two measurement periods. There were smaller, less reliable differences in three of the four ocular measures, in the expected directions, across the two periods. We attempted to use hierarchical stepwise multiple linear regression as a first step in assessing the specificity of the three sensitive ocular measures for predicting the effects of the fatigue manipulation on cognitive performance. We selected simple response time and simple cognitive processing throughput as targets for prediction. Our strategy was to use the first 3 weeks of data to develop the regression equations, and then the last week of data for an assessment of specificity. The fatigue-related changes (from 21:00 to 03:00) in the three ocular measures were such poor predictors of the changes in logical reasoning throughput and simple response speed that we were unable to proceed with the specificity analysis. Overall, we were able to conclude that the individual ocular measures were less sensitive than the performance and subjective measures to this particular fatigue manipulation. However, we were able to draw no useful conclusion with respect to the relative specificity of the combined ocular measures for predicting performance effects.

15. SUBJECT TERMS

Sensitivity, specificity, FIT 2500, oculometer, oculometric measures, fatigue, performance measures, subjective measures, cognitive measures, sleep deprivation, hierarchical stepwise linear regression

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** James C. Miller
a. REPORT Unclass	**b. ABSTRACT** Unclass	**c. THIS PAGE** Unclass	Unclass	19	**19b. TELEPHONE NUMBER** *(include area code)* (210) 536-3596

Table of Contents

Figures

Tables

INTRODUCTION

The objective of this proposed effort was to determine when and how oculometry may be used most appropriately in fatigue research. We proposed to compare the sensitivity of oculometric measures under fatigue stress to the sensitivities of performance and subjective measures, and to assess the specificity of oculometrics for predicting performance effects. The effort was accomplished under the auspices of a planned and funded study that involve oculometry and were conducted in the Fatigue Countermeasures Laboratory at Brooks City-Base, Texas. Our hypothesis (h_1) was that oculometry would be more sensitive and specific for fatigue effects than other physiological, performance and subjective measures used commonly in fatigue research. By oculometry, we meant the measures, saccade velocity, baseline pupil size and pupil response latency as acquired using the FIT 2500 hardware and software (PMI Inc., Rockville, MD). The methods used to collect oculomotor measurements were relatively non-intrusive, suggesting usefulness in some operational settings.

The data comparisons were made as part of the analyses conducted under the reviewed and approved Department of Defense research protocol FBR-2001-15H, "Sleep Aids, Sudden Awakening and Performance." In this study, the hypnotic Zolpidem tartrate (Ambien®) and the naturally occurring hormone melatonin were compared systematically at two doses each for effects on daytime sleep and nighttime cognitive performance and mood in an operationally-relevant paradigm. The subjects worked all night. Subsequently, an Early Sleep Group slept from 0800-1600 and a Late Sleep Group slept from 1400-2200. The subjects worked all night again, and recovery sleep was monitored the following day and night. Measures included polysomnography, oculometry, activity, simple and complex cognitive task performance, vigilance, subjective data, salivary melatonin and cortisol, and vital signs. This report focuses on the oculometry, simple cognitive tasks and subjective reports from the initial night of total sleep deprivation, before the sleep aids were administered.

METHODS

The experiment was structured as a mixed, 2-factor, 2 x 5-level factorial with repeated measures on Factor B. Factor A was Sleep Schedule (early vs. late) and Factor B was Dose (placebo, zolpidem 10 and 20 mg and melatonin 5 and 10 mg).

The subjects consisted of 16 healthy individuals, participating in four groups of four, recruited from the San Antonio area. Males and females between the ages of 18 and 40 years were recruited. All subjects were screened for age and past medical history. Each subject was paid $10 per hour for the 9 hours of training, the 9 hours of baseline sleep and the 302.5 hours of study testing. All candidate subjects were screened to avoid drug interactions. Subjects who acknowledged significant sleeping difficulties were excluded, as well as subjects who admitted to the use of any sleep medication, or medications used in the treatment of narcolepsy or depression. Female subjects who chose to participate in this study were required to submit to a urine pregnancy test within 36 hours prior to the planned experiment. The Subjects were trained on all cognitive tasks during 9 hours of training evenings.

The investigation was carried out in the AFRL Fatigue Countermeasures Lab at Bldg. 1192, Brooks City-Base. This facility was a temporal isolation facility for conducting research and development activities on fatigue countermeasures. Some of its characteristics and capabilities included:

- A 2,100-sq-ft lab with four bedroom-bathroom areas and Annex for meals and entertainment:
    - Continuous subject monitoring through sophisticated audio/video
    - Cognitive performance assessment instruments for individual and/or group performance
    - Physiological measurement capabilities including QEEG, polysomnography, oculometry, body temperature, blood pressure, HR, vestibular function, strength, endocrine levels
- Access to biochemistry laboratory with high pressure liquid chromatography, molecular biology, chemistry, histology, and image analysis systems

Subjects were instructed not to drink alcoholic beverages during the afternoon and evening prior to the scheduled sessions. Caffeinated drinks were not allowed during any test sessions; decaffeinated soft drinks, water, and juice were offered for consumption. Subjects were instructed to go to sleep between 2130 and 2200 hours the night before the scheduled test session, and to awaken between 0600 and 0700 hours. These instructions were intended to reduce variability in the amount of sleep participants obtained prior to the test session.

Subjects were monitored closely to ensure their wakefulness throughout the test sessions. Subjects were not allowed to sleep, doze, or "rest" their eyes at any time during their active participation in the study. They were asked to attempt their best performance at all times during testing sessions.

Oculometry

The FIT 2500 (Figure 1; PMI, Inc., Rockville, MD), was designed originally as an industrial fitness-for-duty evaluation system that would detect physiological impairments due to fatigue and many other factors. It was used here solely as an oculometer, testing involuntary responses. Minimal training was required and no learning or skill effects were expected. Testing took about 30 seconds.

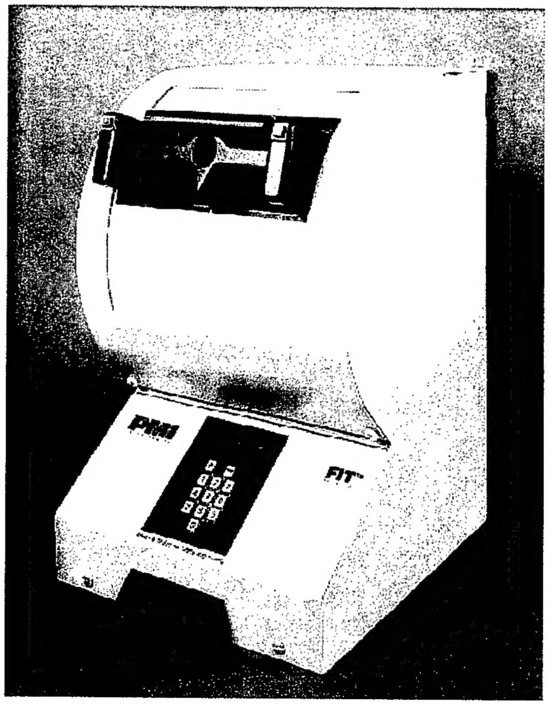

Figure 1. The FIT oculometer.

Saccade velocity slowing may (Russo et al, 1999, Rowland et al., 1997, Stampi et al, 1994;) or may not (Morris and Miller, 1996) be a useful index of fatigue. Baseline pupil size varies as a function of fatigue and or sleepiness (Pressman, DiPhillipo, & Fry, 1986; Schmidt, Jackson, and Knopp, 1981; Yoss, 1969; Ranzijn and Lack, 1997). Similarly, increasing pupil response latency may (Russo et al, 1999, Rowland et al., 1997) or may not (Ranzijn and Lack, 1997) be a useful indicator of fatigue. However, the combining of these three measures, an approach used with the FIT, has allowed reliable detections of fatigue (personal communications, J. Krichmar and R. Perry, PMI, Inc.). We expected to find fatigue effects expressed as reductions in baseline pupil size and saccade velocity, and increased pupil response latency.

Cognitive Tasks

The simple cognitive tasks used here were selected from the Automated Neuropsychological Assessment Metrics (ANAM), a set of standardized batteries designed for clinical use and derived from the Office of Military Performance Assessment Technology's (OMPAT), Tester's Workbench (TWB). The TWB was a library of tests constructed to meet the need for precise measurement of cognitive processing efficiency in a variety of psychological assessment contexts (Reeves et al., 1997). We used the continuous performance, matching to sample, mathematical processing, and logical reasoning tasks. The following descriptions were taken from Reeves et al. (2001).

Running memory (continuous performance) was a continuous letter comparison task (Stanny, 1994). The subjects were asked to monitor a randomized sequence of upper-case letters, A through Z. The letters were presented one at a time in the center of the screen. The subjects were asked to monitor the letters continuously and press a specified key or button if the letter on the screen matched the letter that immediately preceded it. They were requested to press a different response button or key if the letter did not match the immediately preceding letter.

Matching to Sample was a task in which the subject was required to match a block pattern from memory. A single 4 x 4, checkerboard matrix was presented in the center of the screen as a sample stimulus. For each trial presentation of a matrix, the number of cells that were shaded varied at random. Following a specified time interval, two comparison matrices were presented side by side. One of the comparison matrices matched the "sample" matrix, while the other comparison matrix differed in shading from the "sample" by one cell. The subject's task was to indicate, by pressing the appropriate response button, which matrix matched the "sample" matrix.

During the mathematical processing task, arithmetic problems were presented in the middle of the screen. The task involved deducing an answer and then deciding if the answer was greater-than or less-than the number five. Each problem included two mathematical operations (addition and/or subtraction) on sets of three single-digit numbers (e.g., 5 + 3 - 4 = ?). The subject was instructed to read and calculate from left to right and then to indicate whether the answer was greater-than or less-than five by pressing one of two specified response buttons. The operators and operandi were selected at random with the following restrictions: only the digits 1 through 9 were used; the correct answer could be any number from 1 to 9 except 5; greater-than and less-than stimuli were equally probable; cumulative intermediate totals had a positive value; working left to right the same digit could appear twice in the same problem unless it was preceded by the same operator on each occasion (e.g., +3 and +3 were acceptable, while +3 and -3 were not); and the sum of the absolute value of the digits in a problem had to be greater than 5.

The logical reasoning task was an adaptation of the task developed by Baddeley (1968). That task was a linguistic task requiring knowledge of English grammar and syntax. It also required the ability to determine whether various simple sentences correctly described the relational order of two symbols. The present symbolic implementation differed from the original paper and pencil version in that stimulus pairs were presented one at a time and were screen-centered rather than left-justified to reduce differences in visual search times. On each trial, the symbol pair "# &" or "& #" was displayed along with a statement that correctly or incorrectly described the order of the letters as depicted in the example below:

&#

# is first

The subject was required to decide as quickly as possible whether the statement was true or false and then to press the corresponding response button.

For each of these ANAM tasks, the software reported mean response time for correct responses ($MnRT_C$), the standard deviation of these response times ($SDRT_C$), and response throughput (TP). Response throughput was calculated by the ANAM software as:

$$\%\text{Correct x } (60{,}000 / \text{mean } RT_{ALL})$$

where 60,000 = msec/min and RT was in milliseconds.

We also used the Psychomotor Vigilance Test (PVT; Model PVT-192, CWE, Inc., Ardmore PA, available from Ambulatory Monitoring, Inc., Ardsley NY). The following quoted description was provided by Dr. David Dinges (public communication, ca. 2002) and describes the methods we used. The PVT "is a test of behavioral alertness invented by David F. Dinges, Ph.D. and John W. Powell, M.A. It involves a simple (as opposed to choice) portable reaction time (RT) test designed to evaluate the ability to sustain attention and respond in a timely manner to salient signals (Dinges & Powell, 1985). PVT performance has been demonstrated to be highly sensitive to behavioral alertness associated with an interaction of homeostatic sleep drive and circadian phase, total sleep deprivation, cumulative partial sleep loss, slow eyelid closures of the kind experienced by drowsy drivers, etc. [references available].

"The PVT was designed to be simple to perform, to be free of a learning curve (Powell et al., 1999) or influence from acquired skills (aptitude, education), and to be highly sensitive to an attentional process that is fundamental to normal behavioral alertness. The PVT task consists of responding to a small, bright red light stimulus (LED-digital counter) by pressing a response button as soon as the stimulus appears, which stops the stimulus counter and displays the RT in milliseconds for a 1-second period. The inter-stimulus interval varies randomly from 2 seconds to 10 seconds, and the task duration is typically 10 minutes (which yields approximately 80 RTs per trial). The subject is instructed to press the button as soon as each stimulus appears, in order to keep the reaction time as low as possible, but not to press the button too soon (which yields a false start [FS] warning on the display). At the beginning and end of the task a visual analog scale is presented along a subjective dimension of the experimenter's choice."

The variables provided by the PVT-192 included reaction times and their reciprocals, and false responses. It also logged wrong responses (wrong button) and false responses (any button press that occurred when the counter was not running). The task provided a relatively pure demand for sustained, focused attention. It was presented in the recommended 10-minute trial length in the visual-only (0.5-inch LED) mode. The data of interest were the mean of the reciprocals of all reaction times (MRRT; response speed) and the standard deviation of this measure (Dinges et al., 1997).

Subjective Data

Personality, mood and subjective ratings data were acquired from each subject. Subjective ratings of sleepiness and fatigue were acquired using the Stanford Sleepiness Scale (SSS) and the Profile of Mood States. According to Mitler et al. (2000), "Advantages of the SSS include its brevity and ease of administration and the fact that it can be administered repeatedly. Experimentally-induced sleep deprivation increases SSS scores; however, normative data do not exist." The SSS usually correlates with standard measures of performance. However, the extreme values on the scale are used infrequently and the rank-ordered statements overlap several perceptual dimensions including sleepiness-wakefulness, alertness and concentration. Horne (1991) suggested parallelism between the SSS and the alertness-sleepiness descriptors used for the "vigor" factor of the Profile of Mood States (POMS). The POMS vigor scale has also demonstrated sensitivity and reliability with respect to quantifying perceptions of sleepiness.

To use the SSS (Hoddes et al., 1973), the subject selected one of seven sets of Likert-scale descriptors, ranging from 1, "Feeling active and vital; alert; wide awake," to 7, "Almost in reverie; sleep onset soon; lost struggle to remain awake." A rating of 5 or above is often cause for concern with respect to acceptable job performance.

The Profile of Mood States (POMS; McNair et al., 1971) measured dimensions of affect or mood. It consisted of 65 adjectives describing feeling and mood to which the subject responded according to a five-point scale ranging from "Not at all" to "Extremely." Results were reported as six mood factors, of which two were considered here:

- Fatigue-Inertia: weariness, inertia and low energy level
- Vigor-Activity: vigorousness, ebullience, and high energy (mentioned by Horne (1991), as cited, above)

DATA ANALYSES AND RESULTS

Data were available from 14 male and female subjects across the ocular, performance and subjective measures. Within treatment conditions, the data following the 06:00 measurement period on day 2 were discarded. This was the point at which the within-subject treatment (random order of presentation of sleep aid or placebo) and the between-subject treatment (AM or PM sleep group) took effect. Prior to this measurement period, all subjects simply experienced all-night wakefulness under each treatment condition. (While the effects of zolpidem and melatonin on the relative sensitivities and specificities of the measures were of interest, that assessment fell outside the scope of this analysis.)

The observations were then sorted chronologically within subjects, without regard to treatment condition. The chronology included five measurement periods during each of five nights of total sleep deprivation, with at least one week between sleep deprivation nights. The weeks were labeled Week 1 through Week 5. The measurement periods were labeled Period 1 through Period 5. The start times of the measurement periods were 20:00. 21:00, 00:00, 03:00, and 06:00, and each required approximately 30 minutes for all data acquisition.

There were a number of missing FIT data due to the inability of the FIT hardware or software to register the presence of a subject. Out of a maximum of (14 subjects x 5 weeks x 5 periods =) 350 observations, there were 43 (12.3%) missing FIT observations. The missing FIT data were not randomly distributed across subjects and time. One subject (S14) accounted for 11 missing observations and the fifth week accounted for another 19 of the missing observations. This subject and the fifth week were dropped from further consideration.

The resulting genders and ages of the subjects were as listed, below. The mean ages of the two groups did not differ ($p_t = 0.94$).

- 8 males, 21 to 39 yrs (mean 29.1 +/- sd 7.1 yrs)
- 5 females, 21 to 44 yrs (28.8 +/- 8.8 yrs)

Out of the reduced maximum of (13 subjects x 4 weeks x 5 periods =) 260 FIT observations, there were 13 (5.0%) missing. These missing FIT data were relatively randomly distributed across subjects and time, as shown in Table 1. In this latter data set, there were three missing data points for the PVT (S1, Week 1, Periods 1 and 2; S9, Week 1, Period 3) and one missing data point for the Math task (S12, Week 3, Period 5).

Table 1. Distribution of missing FIT data across subjects and time (period number).

S#	N	Week 1	Week 2	Week 3	Week 4	Missing
1	17			P5	P1,P4	3
2	20					0
3	20					0
4	20					0
5	18	P1			P3	2
6	20					0
7	20					0

9	16		P2		P1,P2,P4	4
10	20					0
11	18		P2	P3		2
12	20					0
13	20					0
17	18	P2,P4				2
TOTAL:	*247*	*3*	*2*	*2*	*6*	*13*

As is common in any attempt to compare "apples to oranges," common ground was sought through the process of converting all measures to standard scores. This conversion was accomplished in the usual manner, by subtracting the mean from each observation and dividing the difference by the standard deviation $[\ (x_S - \text{mean}_S)\ /\ \text{sd}_S\ ]$. (Obviously, division by a constant number did not change the distributions.) The general effect of accomplishing the standardization process within subjects instead of across subjects was to eliminate the inter-subject differences among within-subject grand means and to reduce sharply the relative size of inter-subject variability, compared to intra-subject variability, within each measure. The 13 missing FIT data points and 3 missing performance data points were set to zero, the subject mean.

An assumption was made that the fatigue manipulation, *i.e.*, total sleep deprivation, would generate a difference between measure values obtained at 21:00 and 03:00 the next morning. Each of the 4 ocular, 12 ANAM, 2 PVT, and 3 subjective measures was assessed by a 2-tailed, paired t test across the (13 subjects x 4 weeks per subject =) 52 samples acquired at 21:00 and again at 03:00. (The possibility of generating a spurious false positive result across the 21 t tests was not considered to be an important issue at this point in the analysis.)

The differences for all 3 of the subjective measures (SSS and POMS) were statistically significant at $p < 0.001$ (Figure 2). We concluded that the fatigue manipulation was successful in terms of inducing perceptions of fatigue and sleepiness.

Similarly, the differences for 13 of the 14 performance measures (ANAM and PVT) were statistically significant at $p < 0.001$ and the other was significant at $p < 0.01$ (Figure 3). We concluded that the fatigue manipulation was also successful in terms of impairing the subjects' performance of simple cognitive tests.

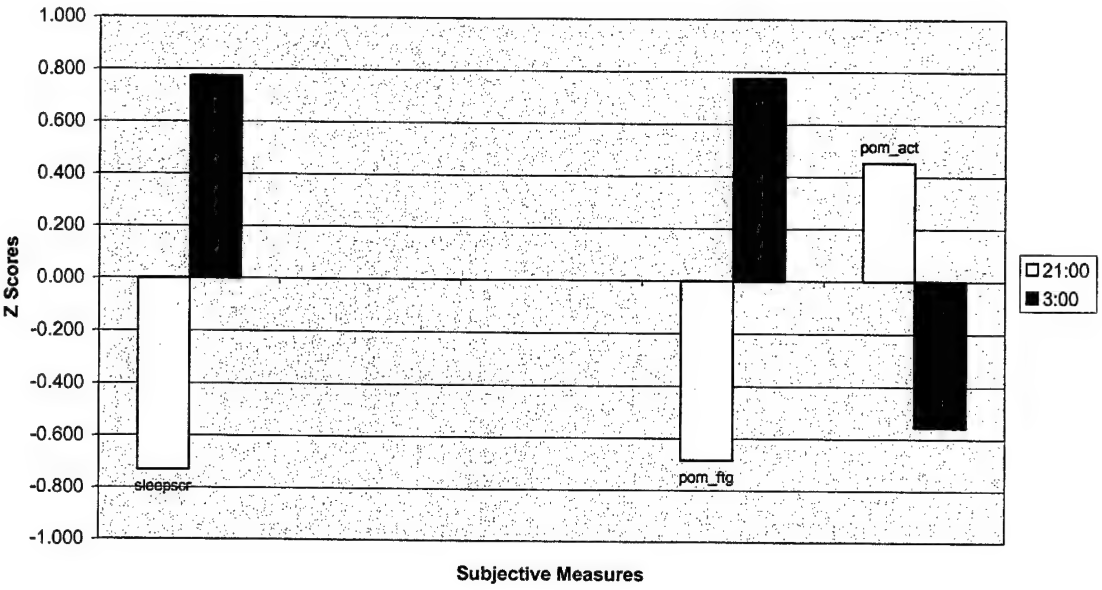

Figure 2. Mean standard scores at 21:00 (light bars, with labels) and 03:00 (dark bars) for the SSS (slpscr) and the POMS Fatigue-Inertia (pom_ftg) and Vigor-Activity (pom_act) factors. All differences significant at p < 0.001. Note that SSS sleepiness and POMS fatigue were greater and POMS activity was lower at 03:00 than at 21:00.

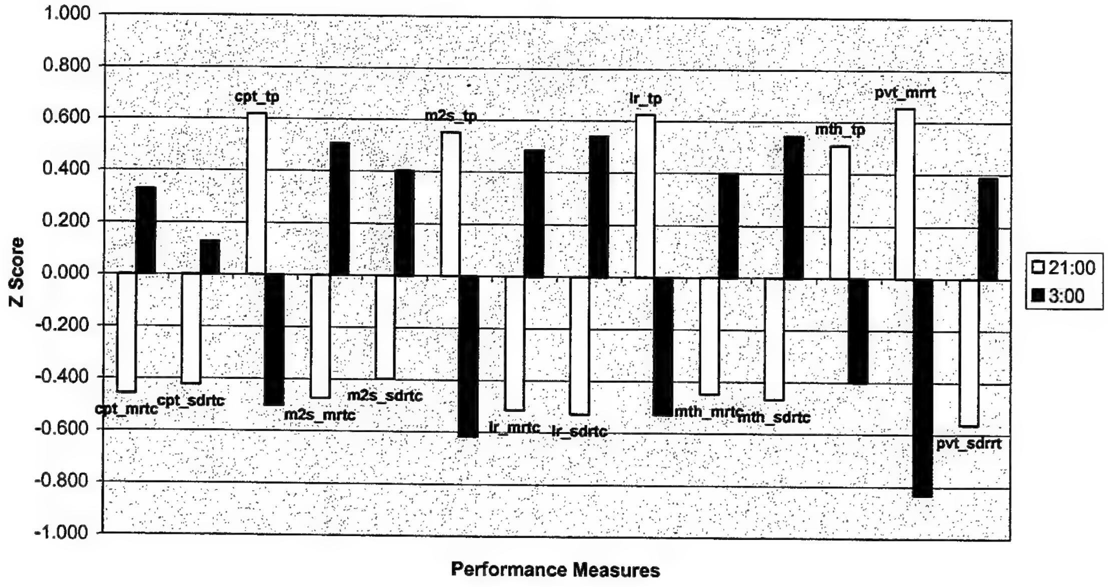

Figure 3. Mean standard scores at 21:00 (light bars, with labels) and 03:00 (dark bars) for the ANAM continuous performance, matching to sample, logical reasoning and math tasks (cpt, m2s, lr, and mth, respectively) measures of response time correct, standard deviation of response time correct and throughput (rtc, sdrtc and tp, respectively); and for the mean reciprocal response time (response speed) and standard deviation thereof for the PVT. All differences significant at p < 0.001 except cpt_sdrtc at p < 0.01. Note that ANAM response times and their variabilities were greater, and throughput was lower, at 03:00 than at 21:00. Also, PVT response speed was slower and its variance was greater at 03:00 than at 21:00.

Generally, the ocular measures did not display differences that were as reliable as the preceding measures when subjected to this comparison (Figure 4). Saccade velocity was slower ($p = 0.001$), pupil constriction amplitude was greater ($p = 0.012$), and resting pupil diameter was smaller ($p = 0.065$) at 03:00 than at 21:00. There was no reliable difference in pupil constriction latency. Thus, we concluded that the ocular measures were less sensitive than the performance and subjective measures to this particular fatigue manipulation. Two ocular measures were significantly affected by fatigue, but, as shown later in this paper, none of the ocular measures was predictive of performance.

The shared variances of the 4 ocular measures with the 14 performance and 3 subjective measures across all 260 observations were assessed as the square of the Pearson product-moment coefficient (r^2). The highest shared variance was 14.0%, between saccade velocity and PVT response speed. Thus, the ocular measures appeared to respond differently to the fatigue manipulation than did the performance and subjective measures.

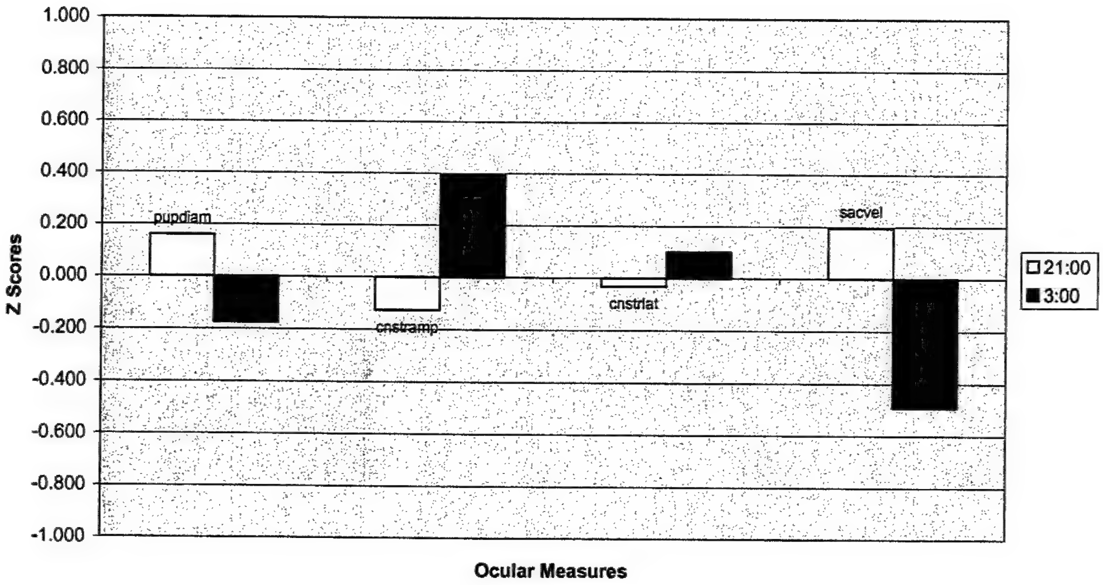

Figure 4. Mean standard scores at 21:00 (light bars, with labels) and 03:00 (dark bars) for resting pupil diameter (pupdiam; $p = 0.065$), pupil constriction amplitude (constramp; $p = 0.012$), pupil constriction latency (constrlat; $p = 0.46$), and saccade velocity (saccvel; $p = 0.001$). Note that resting pupil diameter was smaller, constriction amplitude was greater, and saccade velocity was slower at 03:00 than at 21:00.

We selected hierarchical stepwise multiple linear regression (Cohen and Cohen, 1975) as an initial step approach in assessing the specificity of the three sensitive ocular measures (Figure 3, above) to the effects of the fatigue manipulation on cognitive performance. We implemented the regression analysis in software program BMDP 2R (Release 7.0, BMDP Statistical Software, Inc., 1993). We then selected two performance measures as targets for prediction: simple response time as represented by PVT response speed and simple cognitive processing as represented by throughput in the ANAM logical reasoning task. Both of these performance measures had responded reliably to the fatigue manipulation (Figure 2, above). Specifically, we examined differences in standard (z) scores between the 21:00 test period and the subsequent 03:00 test period.

As a point of comparison for the ocular measures, we also made performance predictions using three subjective measures: the SSS and the two sensitive POMS factors, Fatigue-Inertia and Vigor-Activity (Figure 1, above). We used the same regression approach. To summarize, we ran the following four hierarchical regression analyses on difference scores:

- Y = PVT response speed
    - X = ocular measures
    - X = subjective measures
- Y = logical reasoning throughput
    - X = ocular measures
    - X = subjective measures

The three ocular measures were not strongly inter-related across the whole, 260-observation data set (Pearson r <= 0.179; Table 2). For the ocular measures, based upon their relative sensitivities and reliabilities, we hypothesized that saccade velocity, entering the equation first, would be significantly predictive; that constriction amplitude, entering second, would add significantly to the prediction; and that pupil diameter, entering third, would also add significantly to the prediction.

Table 2. Intercorrelation matrix for ocular measures (260 observations; Pearson r values): pupil diameter, constriction amplitude and saccade velocity.

	Constr Ampl	Sacc Vel
Pup Diam	-0.160	0.179
Constr Ampl		-0.027

For the subjective measures, we noted that the 260-observation correlation between the SSS and the POMS Fatigue-Inertia was r = 0.78 (Table 3; shared variance = 61%). Thus, both were likely to be somewhat equivalent predictors. We opted to use the richness of the POMS factors as our primary predictors, hypothesizing that POMS Fatigue-Inertia, entering the equation first, would be significantly predictive; that POMS Vigor-Activity, entering second, would add significantly to the prediction; and that SSS, entering third, would not add significantly to the prediction because its predictive function would have been accomplished already by the similarly-distributed variance in the POMS factors.

Table 3. Intercorrelation matrix for subjective measures (260 observations; Pearson r values): Stanford Sleepiness Scale (SSS), POMS Fatigue-Inertia (POMS F-I) and POMS Vigor-Activity (POMS V-A).

	POMS F-I	POMS V-A
SSS	0.776	-0.660
POMS F-I		-0.622

As a strategy, we opted to use the first 3 weeks of data (3 weeks x 13 subjects = 39 observations) to develop the regression equations, and then the last week of data for an assessment of specificity. The holding back of a portion of data to be used to test the usefulness of predictive equations is quite common. The unusual aspect of our approach was that each subject was represented three times in the regression database. We were hesitant to reduce the effects of intrasubject variability by combining data across weeks in this applied analysis. We were more interested in learning something about the usefulness of the predictions in daily life, across time, than in understanding the underlying mediation of a phenomenon. As a caveat here, we found that the intra-subject stability for <u>all</u> of the measures was quite poor across these two test periods (Table 4).

Table 4. Grand mean within-subject correlation coefficients (Pearson r; via z transform) across Periods 2 and 4 (21:00 and 03:00, respectively) for Weeks 1 through 4 (52 observations).

				PVT		POMS	POMS
Pup Diam	Constr Ampl	Sacc Vel	LR TP	SPD	SSS	FI	VA
-0.010	0.017	0.160	0.121	0.514	0.179	0.089	0.124

The fatigue-related differences (21:00 minus 03:00) in the three ocular measures from Weeks 1 through 3 were, in fact, very poor predictors of the difference in logical reasoning throughput. The hypothesis that saccade velocity difference would be a good predictor was not supported (F(1, 37) = 0.07), the hypothesis that constriction amplitude difference would be a good second predictor was not supported (F(2, 36) = 0.09) and the hypothesis that pupil diameter difference would be a good third predictor was not supported (F(3, 35) = 0.09). The variance in logical reasoning throughput predicted by the combination of the three ocular measure differences (multiple r^2) was only 0.8%. Thus, this equation was not applied to the data from Week 4.

The fatigue-related differences in the three ocular measures were also poor predictors of the difference in simple response speed. The hypothesis that saccade velocity difference would be a good predictor was not supported (F(1, 37) = 0.10), the hypothesis that constriction amplitude difference would be a good second predictor was not supported (F(2, 36) = 0.74) and the hypothesis that pupil diameter difference would be a good third predictor was not supported (F(3, 35) = 0.51). The variance in simple response speed predicted by the combination of the three ocular measure differences (multiple r^2) was only 4.2%. Thus, this equation was not applied to the data from Week 4.

The fatigue-related differences in the subjective measures were also relatively poor predictors of the difference in logical reasoning throughput. The hypothesis that POMS Fatigue-Inertia difference would be a good predictor was barely supported (F(1, 37) = 2.56, MSe = 0.57, $p = 0.118$), the hypothesis that POMS Vigor-Activity difference would be a good second predictor was not supported (F(2, 36) = 1.30) and the hypothesis that SSS difference would be a good third predictor <u>was</u>, surprisingly, supported (F(3, 35) = 3.26, MSe = 0.50, $p < 0.05$). However, the variance in logical reasoning throughput predicted by the combination of the three subjective measure differences (multiple r^2) was only 21.9%. Thus, this equation was not applied to the data from Week 4.

Similarly, the fatigue-related differences in the subjective measures were poor predictors of the difference in simple response speed. The hypothesis that POMS Fatigue-Inertia difference would be a good predictor was not supported (F(1, 37) = 0.47), the hypothesis that POMS Vigor-Activity difference would be a good second predictor was not supported (F(2, 36) = 0.23) and the hypothesis that SSS difference would be a good third predictor was, again surprisingly, supported (F(3, 35) = 2.31, MSe = 0.50, p = 0.093). However, again, the variance in simple response speed predicted by the combination of the three subjective measure differences (multiple r^2) was only 16.6%. Thus, this equation was not applied to the data from Week 4.

DISCUSSION

To summarize, we wished to compare the sensitivity and specificity of oculometric measures under fatigue stress to performance and subjective measures. We used data from the first night of sleep deprivation in a sleep aids study, when no experimental treatment differences had yet been applied. Each of 13 subjects was represented four times in the final data set, with each of these four nights separated by at least a week. We focused on oculometric, simple cognitive task and subjective data and on two test periods within the night of sleep deprivation, one at 21:00 and a subsequent period at 03:00. All data were standardized as within-subject z scores across weeks. The ocular measures did not correlate well with the performance and subjective measures, suggesting that they responded differently to the fatigue manipulation. There were large, reliable differences in cognitive task performance and subjective assessments, in the expected directions, across the two measurement periods. There were smaller, less reliable differences in three of the four ocular measures, in the expected directions, across the two periods. Thus, we concluded that the ocular measures were less sensitive than the performance and subjective measures to this particular fatigue manipulation.

We then attempted to use hierarchical stepwise multiple linear regression as a first step in assessing the specificity of the three sensitive ocular measures to the effects of the fatigue manipulation on cognitive performance. We selected simple response time and simple cognitive processing throughput as targets for prediction. Our strategy was to use the first 3 weeks of data to develop the regression equations, and then the last week of data for an assessment of specificity. The fatigue-related changes (from 21:00 to 03:00) in the three ocular measures were such poor predictors of the changes in logical reasoning throughput and simple response speed that we were unable to proceed with the specificity analysis.

Our assumption that the fatigue manipulation, *i.e.*, total sleep deprivation, would generate a difference between measure values obtained at 21:00 and 03:00 the next morning seemed reasonable from the viewpoint of fatigue modeling (Sleep Activity, Fatigue and Task Effectiveness (SAFTE) model; Hursh, 1998; Hursh et al., 2004). Our fatigue model predicted that performance would decline from about 94% of expected good performance at 21:00 to about 70% at 03:00 (Table 5).

Table 5. SAFTE model prediction of cognitive performance for the 30-minute periods beginning at 21:00 and 03:00.

Start	Effectiveness
21:00	93.6%
3:00	69.6%

Our conclusion that the fatigue manipulation was successful in terms of inducing perceptions of fatigue and sleepiness and their performance of simple cognitive tests was straightforward, considering the sizes (usually greater than one standard deviation unit) and reliabilities (p < 0.001) of the differences in the means. Our conclusion that the ocular measures were less sensitive than the performance and subjective measures to this particular fatigue manipulation was also straightforward considering the sizes (much less than one standard deviation unit) and reliabilities (p = 0.065 to 0.001) of the differences in the means.

In a similarly structured analysis of nocturnal saccade velocity, using a more sophisticated oculometer in another study[1], there was "a significant main effect of Trial [similar evening and early morning times] on peak saccadic velocity that appeared to be fatigue related (F(1,11) = 13.08, MSe = 1387.0, p = 0.004). Peak saccadic velocity was significantly faster during Trial 1 than during Trial 2. ...there was a similar, significant main effect of Trial on mean saccadic velocity that appeared to be fatigue related (F(1,11) = 18.74, MSe = 358.1, p = 0.001). Mean saccadic velocity was significantly faster during Trial 1 than during Trial 2." These effects are shown in Figures 5 and 6, taken from the cited draft report. Thus, we suspect that the FIT oculometer, in the version that existed in the fall of 2001, did not measure saccade velocity adequately.

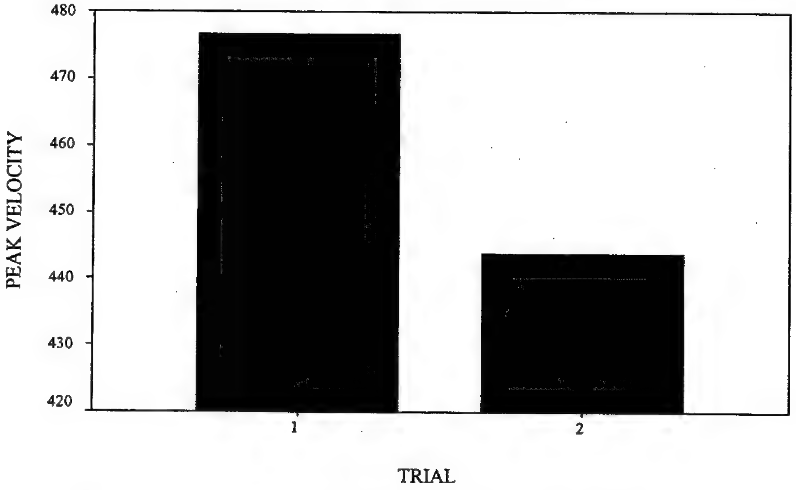

Figure 5. Main effect of Trial on peak saccadic velocity.

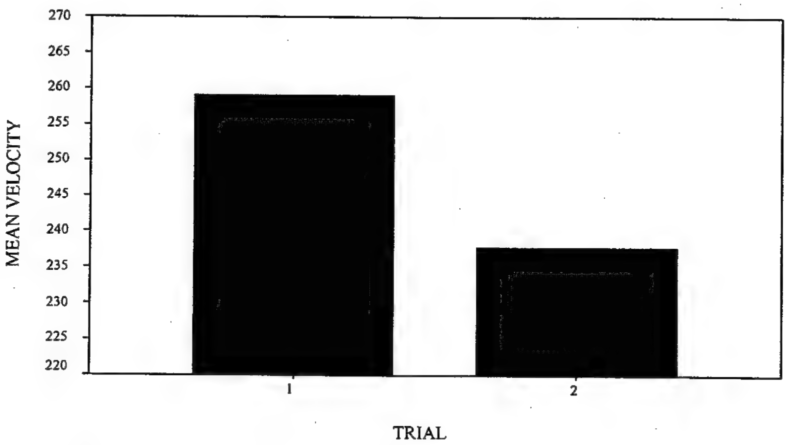

Figure 6. Main effect of Trial on mean saccadic velocity.

[1] Eddy D et al. *An Assessment Of Modafinil for Vestibular and Aviation-Related Effects*, USAF Surgeon General research protocol no. FBR-2000-35H (technical report in preparation).

Our strategy of using the first 3 weeks of data to develop the regression equations, and then the last week of data for an assessment of specificity was straightforward. Holding back a portion of data to be used to test the usefulness of predictive equations is quite common. The unusual aspect of our approach was that each subject was represented three times in the regression database. We were hesitant to reduce the effects of intrasubject variability by combining data across weeks in this applied analysis. We were more interested in learning something about the usefulness of the predictions in daily life, across time, than in understanding the underlying mediation of a phenomenon.

In fact, the fatigue-related differences in the three ocular measures were very poor predictors of the difference in logical reasoning throughput and simple response speed. However, we found that the intrasubject stability of all of the measures was quite poor across the two test periods. This was probably the quantitative cause of our inability to predict simple cognitive performance using the oculometric data.

Overall, we were able to conclude that the individual ocular measures were less sensitive than the performance and subjective measures to this particular fatigue manipulation. However, we were able to draw no useful conclusion with respect to the relative specificity of the combined ocular measures for predicting performance effects.

ACKNOWLEDGMENTS

This project was conducted under AFRL Job Order Numbers (JON) 71845901 and R3ARMEEG, the latter funded by the Department of Defense Saccadic Fatigue Measurement Research Program. The experimental protocol was reviewed for subject protection by the Brooks City-BaseInstitutional Review Board and the AF Surgeon General's Research Oversight Committee and assigned AF/SG protocol no. FBR-2001-015H. The Ancillary and Clinical Pharmacy Flight of the Air Force Wilford Hall Medical Center prepared the medications and supported our double-blinding procedures. Mr. Patrick Hickey of NTI, Inc., was responsible for data reduction and management.

REFERENCES

Cohen J, Cohen P (1975). *Applied Multiple Regression/Correlation for the Behavioral Sciences*. Lawrence Erlbaum, Hillsdale, New Jersey.

Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, Aptowicz C, Pack AI (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. *Sleep*, 20(4), 267-277.

Dinges DF, Powell JW(1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. *Behavior Research Methods, Instruments and Computers* 17:652-655.

Hoddes E, Zarcone VP, Smythe H, Phillips R, Dement WC (1973). Quantification of sleepiness: A new approach. *Psychophysiology,* 10, 431-436.

Horne JA (1991). Dimensions to sleepiness. Chapter 7 in TH Monk, *Sleep, Sleepiness and Performance*, Wiley, pp. 169-196.

Horne JA, Ostberg O (1976). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *Int J Chronobiol*, 4(2), 97-110.

Hursh SR (1998). *Modeling Sleep and Performance within the Integrated Unit Simulation System (IUSS)*. Technical Report Natick/TR-98/026L. Science and Technology Directorate, Natick Research, Development and Engineering Center, United States Army Soldier Systems Command, Natick, Massachusetts 01760-5020.

Hursh SR, Redmond DP, Johnson ML, Thorne DR, Belenky G, Balkin TJ, Storm WF, Miller JC, Eddy DR (2004). Fatigue models for applied research in warfighting. *Aviation, Space and Environmental Medicine*, in press.

McNair DM, Lorr M, Droppleman LF (1971), *Manual for the Profile of Mood States*, Educational and Industrial Testing Service (EdITS), San Diego CA. This company is now at: http://www.edits.net/.

Mitler MM, Carskadon MA, Hirshkowitz M (2000). Evaluating sleepiness. Chapter 104 in MH Kryger, T Roth, WC Dement (ed.), *Principles and Practices of Sleep Medicine (3rd ed.)*, WB Saunders, Philadelphia.

Morris TL, Miller JC (1996). Electrooculographic and performance indices of fatigue during simulated flight. *Biological Psychology*, 42, 343-360.

Powell NB, Riley RW, Schechtman KB, Blumen MB, Dinges DF, Guilleminault CA (1999). Comparative model: Reaction time performance in sleep-disordered breathing versus alcohol-impaired controls. *The Laryngoscope*, 109: 1648-1654.

Pressman MR, DiPhillipo MA, Fry JM (1986). Senile miosis: the possible contribution of disordered sleep and daytime sleepiness. *J. Gerontology*, 41, 629-634.

Ranzijn R, Lack L (1997). The pupillary light reflex cannot be used to measure sleepiness. *Psychophysiology*, 34, 17-22.

Reeves D, Kane R, Winter K (1997). *ANAM V3.11a/96 User's Manual: Clinical and Neurotoxicology Subsets.* National Cognitive Recovery Foundation Special Report NCRF-SR-97-01

Reeves D, Winter K, Kane R, Elsmore T, Bleiberg J (2001). *ANAM 2001 User's Manual: Clinical & Research Modules*. National Cognitive Recovery Foundation Special Report NCRF-SR-2001-1.

Rowland L, Krichmar J, Sing H, Thomas M, Thorne D (1997). Pupil dynamics and eye movements as indicators of fatigue and sleepiness." *Proc. Annual Meeting of the Association for Research in Vision and Ophthalmology*, Ft Lauderdale FL.

Russo M, Thomas M, Sing H, Thorne D, Balkin T, Wesensten N, Redmond D, Welsh A, Rowland L, Johnson D, Cephus R, Hall S, Krichmar J, Belenky G (1999). Saccadic velocity and pupil constriction latency are sensitive to partial sleep deprivation, and deprivation delated changes correlate with simulated motor vehicle crashes," *American Academy of Neurology Annual Meeting*, Walter Reed Army Institute of Research, Washington DC

Schmidt HS, Jackson EI, Knopp W (1981). Electronic pupillography (EPG): objective assessment of sleepiness and differentiation of disorders of excessive somnolence. *Sleep Res.*, 10, 48.

Stampi C, Aguirre A, Macchi M, Hashimoto S (1994). *Evaluation of Pulse FIT Parameters for Detection of Fatigue (Reduced Alertness),* Institute for Circadian Physiology, Cambridge MA.

Stanny RR (1994). *Effects of sustained performance on cognition and event related potentials*. Unpublished manuscript, Naval Aerospace Medical Research Laboratory.